DOCUMENT RESUME

ED 129 862 TM 005 634

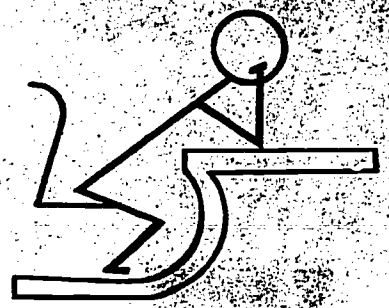
AUTHOR Savage, Edward R.
TITLE A Layman's Guide to the Measurement of Educational Achievement in New Jersey.
INSTITUTION Greater Newark Urban Coalition, N.J.
PUB DATE Mar 76
NOTE 44p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Achievement Tests; *Aptitude Tests; Criterion Referenced Tests; Educational Assessment; Guides; Norm Referenced Tests; Scores; *Standardized Tests; *State Programs; *Testing Problems; *Test Interpretation
IDENTIFIERS New Jersey; New Jersey Educational Assessment Program

ABSTRACT

This paper addresses itself to the history and use of standardized aptitude and achievement tests for practical judgements about schools in New Jersey. In the author's opinion, school-wide aptitude testing is not a particularly useful practice when other types of information, particularly grades, achievement tests, and teacher's comments, are available, though it may be useful when there is no information about the student or when special aptitudes for special programs, are being sought. Whereas aptitude tests claim to be able to measure an individual's future performance, achievement tests are closely tied to things which people have done, to scientifically verifiable accomplishments. Norm-referenced tests classify children on the basis of their relative accomplishments, but have limited value for identifying educational deficiencies as a means of providing effective remediation. Criterion referenced tests strive to overcome that problem. In the final section, the New Jersey State Assessment Program is discussed; the appendixes include a glossary of terms and a guide to test scores. (Author/BW)

NEW JERSEY EDUCATION REFORM PROJECT



Sponsored by:

Greater Newark Urban Coalition, and

A Report of

The New Jersey Education Reform Project

A LAYMAN'S GUIDE TO THE MEASUREMENT OF EDUCATIONAL ACHIEVEMENT IN NEW JERSEY

by
Edward R. Savage
Teachers College
Columbia University

March 1976

TABLE OF CONTENTS

I. INTRODUCTION

Foreign visitors to America have often remarked that our passion for believing in multiple-choice test scores is one of the most inexplicable aspects of our cultural mythology. We use tests everywhere: in business to screen employees, in education to screen kindergarteners, in management to select the "promotables." We even have popular magazine columns which feature tests for recreation. We publish and use over 90% of the standardized tests in the English-speaking world.[1] Many Americans are tested every time they change grades in school and, at some levels, every time they change jobs. A great deal has been written about the American ways of using and misusing standardized testing.

The problems of testing are not confined to schools. Harry Levinson, a noted management consultant, finds that tests are poorly used in business by personnel workers. Seldom is a personnel worker qualified to select and apply tests properly: "At best, his knowledge is a pastiche of concepts. The limitations of his knowledge are made explicit by the wide range of useless psychological tests he buys by the thousands of copies."[2] Before a job interview, for example, an applicant often takes a battery of aptitude and attitude tests, some of which are very general and others closely related to the job. How might the decision to hire be made? The tests, which perhaps took longer than the interview, may merely eliminate the clearly unqualified applicants. The personnel worker, in all likelihood, will then look at high school locations, previous work records, and, with his interview impressions in mind, make a decision. The tests and surveys may just provide a veneer of excuses in such cases for a judgment that is basically a matter of one person's subjective evaluation of another person.

Judgment, then, is the ultimate reason for standardized testing. If a person is to judge another person, there must be democratic criteria, so the American myth goes, because the judgment will be "personal" and subjective otherwise. If judgments about a lot of people are to be made, according to another popular myth, there must be scientific backing for them. But in practice, we may find

---

[1]Oscar Buros, Mental Measurements Yearbook, Seventh Edition (vol. 1), Highland Park, N.J.: Gryphon Press (1972), p. xxxvii.

[2]Harry Levinson, The Exceptional Executive, New York: Mentor Books (1971), p. 262.

that judgments are made personally, and good criteria or unscientific criteria or none, such judgments take too much time to correct and so they stick. Tests often crystallize what we would do or feel anyway. You may take a Reader's Digest vocabulary test monthly. Does your score affect how you feel about yourself, how you judge your ability to persuade other people? Would it affect your community feeling if you knew your neighbor's score? Your working relations if you knew the bosses'? But, on the other hand, what if people only had time to know you by your scores? This gets us into the central issues of standardized testing in democratic schools for the masses. If you are in charge of thirty small persons and have no time to get to know many of them on a personal level before you start judging them, how much help can you get from standardized tests? Once you know the kids, how much do you need those scores? And if you're responsible for 300, 3,000 or 300,000, how important are test criteria for your judgment, in comparison to the economic and social "facts of life?"

This paper, by an administrator rather than a test expert, will address itself to the history and use of standardized aptitude and achievement tests for practical judgments about schools in New Jersey. The recently developed New Jersey State Educational Assessment program will be scrutinized quite closely. The basic argument of the pages that follow is that the generally uncritical use of national tests is not a good way for either classroom teachers or local officials to make judgments about their schools' programs. It is also this writer's point of view that the New Jersey statewide assessment program is generally a better program of educational evaluation for statewide decision making and some local decisions than any of the commercially available and popular standardized test packages.

A little historical analogy may help to set standardized testing into the perspective of decisions about students and about schools. The time-tested, good ol' way of finding out how a student was doing was to have the child stand before the teacher and recite his lesson, whether that be "Horatio at the Bridge" or the ABC's. This method was used before our English forefathers set foot in America, and with a little updating, it might be seen hanging on in the "Friday Testday" syndrome characteristic of the schools even a generation ago. At the level of making judgments about individuals, a personal "diagnosis" of this sort is speeded up by some kinds of standardized testing, and the basic goal of finding out what students have learned from their teachers is still to enable the teachers to teach

the kids what they need to know next.

Another golden practice was the "inspection" visit. Periodically, the School Trustees would visit and hear lessons from the "scholars." On the basis of the students' performances of their 3 R's, the officials could make judgments about teachers, schools, and methods of teaching--and take appropriate financial and punitive actions. To a large extent nowadays, such officials rely on the figures from standardized testing and the budget book instead of their ears and eyes to make judgments about the district's programs. But the old inspection visit had its merits as well as its defects.

At least, real adults met real children, and there were fairly clear standards: either the child could recite his lesson or he couldn't. Some judgment could also be made of the class as a whole by picking various students to speak, and the teacher's relationship with the whole group could be seen.

One defect of this method was that it was conducted in an atmosphere perhaps reminiscent of the Terror of the French Revolution. Today we would call this "test pressure." A second fault is that only a few children would be heard in many cases, and in large cities only a few classes might be visited in a school. Finally, only a smattering of the lessons would be heard. While judgments could be made on these grounds, they could not be "good" judgments because all the facts were never in. This is, in this writer's opinion, the major problem of testing of any sort: how can the information be provided which is necessary to make good judgments about the educational system, the school, and the individual student?

In the first place, different kinds of information are needed for each of these levels of decisions. "Properly used, evaluation should enable teachers to make marked improvements in their students' learning," states one of the major modern texts on classroom testing.[3] To do this, a teacher must know how much of the course a student needs to learn and how he can learn it best. Administrators must provide the books, teachers, and backup to allow this learning to happen. To do this, they have to know what kinds of groups of people will inhabit their schools and districts. And ultimately, Federal and State Department

---

[3]Benjamin Bloom, et al., Handbook of Formative and Summative Evaluation of Student Learning, New York: McGraw-Hill Book Co. (1971), p. v.

of Education officials must know what kinds of problems are not being licked on the lower levels and what kinds of research, tax money, and services can be used to attack such problems.

In the second place, many levels of local decisions rely, in districts using standardized testing, upon the single administration of a mass-testing program during one period of the year. The data are derived from two to three days of testing sessions (6 hours or more total test-taking time). While such information is perhaps useful for building-level judgments, such decisions might equally well be made with less testing time; system-wide decisions on pupils and programs require even less pupil testing data and hence even less testing time than building level decisions. Finally, although mass testing is valuable as a screening device, it has limited utility for developing really meaningful teacher judgments regarding the needs of individual children, for various technical reasons to be discussed in the following sections. For some children, two hours at least of individual testing by a certified psychologist may be needed to prepare school officials to make reasonably careful judgments. Many current school testing programs thus are both more extensive and expensive than organizational information needs require and at the same time less accurate than pupil guidance programs need.

The standardized test is presently the most commonly used method of providing information for classroom, district, state, and federal judgments. In fact, two types of standardized tests, aptitude and achievement, have commonly been used to the exclusion of all other types of data gathering for local decision making until the 1970's. Briefly, an achievement test is supposed to show what students have learned from a particular school program; an aptitude test is supposed to find out how well the student can do a series of harder and harder problems like (but not too much like) the things he will have to do in school in the future. A third type of test has been popularized recently, an achievement test with "criterion-referenced" scores. Subsequent sections of this paper will explain how each type of test can be used most effectively.

II. APTITUDE TESTING

In the previous section, testing was related to the kinds of judgments that people have traditionally made about schools. The chief advantages of standardized tests are that they are administered under uniform conditions, are convenient and easy to use, and information for decisions can be collected quickly. The questions of what kinds of decisions and what types of information will be discussed in the sections on Aptitude Testing, Achievement Testing, and State Assessment Testing. Basically, it will be argued that the less general testing there is, the better it will be not only for children but for decisions at every level. In this section, New York City's banning of "intelligence" tests will be reviewed in order to argue that such testing can be dropped from general testing programs. Standardized aptitude tests may be useful when there is no information about the student or when special aptitudes for special programs are being sought (as in vocational or special educational programs).

* * * *

Bumps on the head, skull measurements, and parental achievements were various means of measuring aptitudes before the modern era of testing was opened by Alfred Binet's work with the retarded in Paris around the turn of the century. Binet's work formed the foundation for individualized aptitude testing by qualified psychologists, an area which will not be discussed in this paper. A second great breakthrough occurred in World War I when Arthur Otis adapted Binet's work to mass testing for the Armed Services.[4] While there have been revolutions in tests and in theory since then, the purposes and methods of mass aptitude testing seem to have changed little. The tests are still designed to enable persons of high aptitude or low aptitude to be singled out for special attention in schools.

Aptitude tests in schools are generally given in conjunction with the achievement testing program. Tests are administered either in the fall or the spring and the scores become part of the students' cumulative folders. The test scores are most often used for making decisions

---

[4]W. James Popham, Criterion-Referenced Measurement: An Introduction Englewood Cliffs: Educational Technology Publications (1971).

about students' educational programs in the spring when guidance counsellors or administrators consult the aptitude scores along with current grades, teacher recommendations, and achievement test scores. Officials give various levels of importance to these categories of student records, teachers' comments often being most important. Students' retention and placement within the classes in a particular grade are the major decisions made on the basis of such data. Let us examine what the aptitude tests are.

An aptitude test is basically a series of increasingly difficult exercises. Those who can only do the easy exercises are said to have a low aptitude; those who can do even the hard ones are said to have high aptitudes. Psychologists prefer to use the term "aptitude" instead of the earlier idea of "intelligence", and for testing purposes, some psychologists now mean by "intelligence" only performance on a test.[5] Naturally, this definition is not what most people mean by "intelligence", and many of the problems raised by testing originate not with the test but the words used to describe and interpret them. People confuse someone's general intelligence with test results. Some parents may feel their child's chances for success in life can be predicted by a two-hour test administered in third and fifth grade. Other parents may argue, more wisely, that their child's life in school may be affected for good or for ill by the way educators use aptitude test scores. Some teachers and administrators may look at aptitude scores and decide who the fast and slow students will be. In such cases, test experts complain that their tests are being misused by parents and educators. Decisions about class groups, for instance, are better made if achievement scores, grades, and teacher recommendations are used with aptitude scores. A review of some characteristics of aptitude tests may suggest why it is sometimes tempting to misuse them.

---

[5] Op.cit., Benjamin Bloom, p. 24. Also, Robert L. Thorndike, Educational Measurement, Washington, D.C.: American Council on Education (1971), p. 545.
An opposing definition of intelligence by D. Wechsler, author of some major individual tests like the Wechsler Intelligence Scale for Children (WISC) indicates how broad and hard to test any other definition would be: "the aggregate capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment."
David Wechsler, "The IQ is an Intelligent Test", New York Times Magazine (June 26, 1966),p. 13.

Of the many types of aptitude tests available, the Otis-Lennon (a revision of Otis's original materials for children) and the Lorge-Thorndike are two of the most popular for use with large groups of children. The Otis contains a series of problems with words and figures and results in a single score. In the Lorge-Thorndike, the word and figure problems are separated and scored separately. Then a combined score like the Otis's can be derived.

The word problems may consist of synonyms or of word puzzles like analogies (tree is to lake as green is to _____?). Short reading passages and questions may also be included. These types of word problems are grouped together and called "verbal aptitude" subtests. The figure exercises may contain arithmetic and geometry problems ranging from simple computation to elaborate problems (including puzzles like mazes and counting boxes in a picture). Depending on the test, these figure problems are called "non-verbal" or "quantitative" aptitude subtests. In terms of student time and information, a counsellor can get a single IQ score using the Otis test in a half hour; the Lorge-Thorndike usually would take twice as long. The total scores from the tests can be used for most of the kinds of judgments that test-givers want to make, or they can be converted into percentiles or some other score which can be compared to other tests more easily. If test makers had their choice of doing it all over again, they might want to stop at this point, for both the simple scores and percentiles are clear and useful.

(The interested reader may refer to the Glossary and Guide to Scores for further details on the utility of various scores.)

Unfortunately, the history of testing took another direction. Lewis Terman devised a ratio between test score and age, and in 1916 he called this ratio "Intelligence Quotient", or IQ. His notion, which is not popular among test experts nowadays, was that IQ would remain constant throughout life; but the idea caught on among the general public. Because there is some general demand for such scores, aptitude tests continue to report IQ's; and many people believe that the IQ figures represent a fixed quantity of intelligence which can be measured once and for all on a test. This number, recall, can be used to label children after no more than two hours of testing on reading, mathematics and pictures. When misunderstanding of these scores is used as the basis for judgments, not only this particular test but the purposes of aptitude testing in general are being distorted.[6]

---

[6]Norman E. Gronlund, Measurement and Evaluation in Teaching, New York: Macmillan Co. (1971), p. 320.

The valid purpose of aptitude testing, under specific conditions, is the selection of persons who are likely to succeed or fail at certain tasks. When employers want to select the person most likely to succeed in a particular job, aptitude tests related to that job are generally helpful in predicting who these will be. Inappropriate aptitude tests can be worse than useless in such cases. Howard Lyman discusses an anecdote about a personnel manager who promoted a young, inexperienced accountant over an older and experienced man because the older fellow's scores on a "spatial relations" subtest of an aptitude battery was lower than that of other job candidates. Spatial relations (often counting the blocks in a printed figure), however, is quite unrelated to the types of work done by accountants!

Schools use test batteries like the Differential Aptitude Tests with teen-agers to try to give the students some idea of the variety of things they can do well. The D.A.T. tests range from verbal/mathematics subtests to clerical speed and mechanical reasoning, and their large variety of subtests make them much more useful for vocational counselling than less complex tests like the Lorge-Thorndike or Otis tests. The U.S. Employment Service uses a similar aptitude test, the GATB, General Aptitude Test Battery. College Boards are specialized occupational tests, designed to show who is likely to succeed during their first year of college.

When the armed forces in the 1950's wanted to offer enlarged opportunities to high school dropouts, aptitude tests of various sorts were essential to selecting the best candidates.[7] But other kinds of records also allow such judgments to be made. School grades also show pretty well what freshman year college grades will be;[8] past job performance predicts very well how well a person will do on similar jobs. When such information is available, in fact, it is dangerous to neglect it.

Lyman relates that a woman was almost fired from a job after the personnel manager gave some tests. The personnel manager knew the facts but, mistakenly, believed the test: "The tests really opened our eyes about her.

---

[7]Ivar Berg, Education and Jobs: The Great Training Robbery, Boston: Beacon Press (1971), pp. 153-54.

[8]Arvo E. Juola, "Prediction of Successive Terms' Performance," American Educational Research Journal (1966), pp. 191-97.

Why, she's worked here for several years, does good work, gets along very well with the others. That test shows how she had us fooled!"[9] When information is not available, aptitude tests can prove useful in making some educated guesses. When a student arrives in New Jersey from the New York City Schools, for example, it may be quite a while before the official transcripts are available. Meanwhile, the administrator must place the student in a class and the teacher must begin an instructional program, and a quick session with the Otis helps provide a better instructional placement of the student.

School-wide aptitude testing is not a particularly useful practice when other types of information, particularly grades, achievement tests, and teachers' comments, are available. Nonverbal (quantitative) aptitude scores tend to line up with math grades and math achievement test scores; the verbal subtests show pretty much the same picture as reading achievement test scores and basal reader progress. Besides, when all this data is collected in a folder, the chances are greater that someone is going to misinterpret some of it; and because of the popularity of the IQ, the aptitude test scores are more likely to be misunderstood than grades or achievement scores. New York City's experience has shown that it is quite possible to do without standardized intelligence tests.

In March, 1964, the New York City Board of Education announced the elimination of group IQ testing in the schools and the substitution of achievement testing plus special aptitude measures for first grade and selective high schools. Popular commentators such as Fred Hechinger hailed this as a much needed move.[10] Even opponents of the test ban admitted that there seemed to be dangers of misusing general aptitude scores.[11] David Wechsler reviewed the issue a couple of years later, after the new first grade testing materials were available, and came to the conclusion that IQ tests had done a better job of estimating school aptitudes,[12] but Wechsler's criticism is not a convincing argument for renewed general aptitude testing. Since then,

---

[9]Howard B. Lyman, Test Scores and What They Mean Englewood Cliffs: Prentice Hall, Inc. (1971).

[10]Fred M. Hechinger, "I.Q. Test Ban," New York Times (March 8, 1964).

[11]Fred M. Hechinger, "More Pros and Cons on Value of I.Q. Tests," New York Times (February 6, 1966).

[12]Op.cit., David Wechsler, p. 12 ff.

repeated evaluations of basic skills learning in New York City have been publicized, none emphasizing IQ or other sociological mitigating factors. It seems to this writer that eliminating the IQ argument also shifted the burden of accountability for students' achievement to the schools and that with reasonable financial support (until 1975) New York City's schools were responding with some success to the challenge to demonstrate improved reading achievement.[13]

To summarize briefly, the argument is: 1. that mass aptitude tests provide information that is generally misunderstood as showing that a fixed IQ number can be attached to every student for his educational career (intelligence, as measured by tests, can change from test to test); 2. that other information such as grades and achievement test scores provide useful information for making basic decisions about students; 3. that aptitude tests are most useful for the special problems that transfer, vocational, or special education students present.

[13]Not an overstated proposition: the AFT-negotiated regulations governing the tested population make it impossible to claim that the improvements are uniform. This example is not being used as a "T&E" model, but merely a case to show that alternatives to intelligence testing are available to New Jersey.

III. ACHIEVEMENT TESTING

The argument thus far has been that standardized testing is basically a series of "mental" tools to aid people in making judgments about other people. As with a hand tool, the problems of testing are more often caused by the way tests are commonly used, not by the way they are designed. In particular, aptitude tests should have very limited application, for the information they produce is most easily misunderstood; and, as the example of New York City's IQ test ban suggests, a combination of school records and other kinds of tests make it possible to make general judgments about grouping students without resorting to mass "intelligence" testing. The section on Achievement Testing begins with a quick review of general achievement tests and delves into the problems of achievement testing. It will be argued that a new kind of achievement test score, the criterion-referenced score, will prove more useful for teachers and parents than "grade-level" scores. This discussion also is intended to prepare the reader for the last section, where the New Jersey State Educational Assessment Program is discussed as one kind of "criterion-referenced" test.

* * * *

Historically speaking mass achievement testing preceded the World War I IQ tests by more than a decade. This history not only illuminates the differences between achievement and aptitude tests, but it also forms an interesting sideline of the movement to apply sound business practices to education.

Just after the turn of the century, Edward L. Thorndike investigated the problem of setting clear, scientific "standards" for learning. Such investigations in industries had been made by efficiency experts like Taylor and Gilbreth. Thorndike and his associates devised "proficiency scales" with which certain school accomplishments could be measured, particularly in handwriting and arithmetic.[14] The idea of making sure that students measured up to a criterion of proficiency became very popular.

In 1911, the New Jersey Legislature initiated statewide testing for high school entrance. Until about 1930 this program of statewide testing in arithmetic, English, and social studies was administered to students seeking to go to high schools, and, occasionally, to high school graduates.[15] This idea lost its popularity about 1930, and the tests were dropped.

---

[14]Raymond E. Callahan, Education and the Cult of Efficiency, Chicago: University of Chicago Press (1962), pp. 100-101

[15]Roy H. Wager, "A,B,C, or None of the Above," NJEA Instruction (1974).

Whereas the aptitude (or IQ) tests which were discussed in the previous section claim to be able to measure an individual's "intelligence" or future performance, achievement tests are closely tied to things which people have done, to scientifically verifiable accomplishments. This process of scientific verification of achievements can be briefly summarized.

Achievement test developers are often teams of university scholars and publishers. Therefore, the most popular tests are named for the university at which the groundwork of test design and research was laid: Stanford, Iowa, and California have been such centers. Other tests are named for the professional groups which utilize their results: the Metropolitan Achievement Tests (developed for New York City), the Co-ops (prepared for independent schools) and the College Boards are such test batteries.

The test development process starts, much as did E.L. Thorndike, by tying the tests as closely as possible to school subjects. Research committees collect course guides for the various subjects from schools and professional associations all over the country. They research state curricula and collections of tests. The developers expend a lot of effort in finding out what topics are taught to just about everybody in the country. When these general topics are identified for a subject area, then often graduate students will prepare groups of test items which, in theory, will show what students will be called upon to do as they pass from grade to grade in American schools. The major scholarly problems are to ensure that the tests get progressively more difficult from level to level and that the tests measure what they say they intend to measure. A pilot version, prepared by the university developers, is often tested nationwide by a publisher to be sure that these two criteria are met. This pilot testing often involves more than 10,000 students, selected as a "sample" of the various groups in the nation. The pilot test scores are used in two ways. First, the test itself is statistically evaluated to be sure that the questions are not too easy or too hard for the students. Questions that are too easy are bad because they don't show who has learned more than most other students in his class. If there are still "bad" items in the test, then either everyone in the grade level will know the skill already or so few students will know the answer that it would be a waste of time to include the question in the test. Such items are either dropped or moved to higher or lower testing levels. The second useful aspect of pilot testing is that the scores can be used to make up national standards or "norms." Because individual and group scores are expressed in relation to these "norms",

such tests are frequently referred to as "norm-referenced" tests. The scoring guides, like aptitude scores, can be set up in various ways. Since achievement tests are broken down into subject areas, there are separate scores for each subject. A total score is also computed for the whole test. The scores are expressed in various kinds of figures. Test experts prefer to use percentiles or related scores like stanines. School officials and teachers often prefer to use a figure called a "grade-equivalent score" which is derived from the analysis of various groups' test scores. Because the percentile and grade-equivalent scores are often taken to mean very different things, many of the problems with misusing achievement tests occur at this point. The Appendix indicates how scores are developed. Here we will be concerned only with how they are used.

The following is one hypothetical rationale for a local school testing program. In most schools, the testing program is scheduled to occur at a particular time of year, commonly in either the mid-spring or mid-fall. These times are chosen for the whole system so that the gradquivalent scores for all grades will be as close as possible in terms of months of the year and the various decisions can be made about students and programs on the basis of consistent data gathered at the same time for everyone year by year. Commonly, a mid-fall testing allows students to complete early reviews of their previous work and provides scores about six weeks after testing which can be used just before Christmas to plan the rest of the year's work. A mid-spring testing produces scores just on time for a final evaluation of the program for the year and for making decisions about planning student's schedules for the next year.

This rationale is shot through with technical and administrative defects. The technical defects revolve around the issues of interpreting scores properly, and these issues are discussed in a little more depth in the Appendix. Here, let it be noted that because test scores are developed by statistical operations and because tests are given en masse, individuals' scores have limited utility for judging any particular student's actual achievement. For example, an individual's score on a particular test may vary from time to time due to the test circumstances, or his physical and emotional conditions. Further, grade-equivalent scores are subject to similar kinds of serious misunderstanding that make IQ scores dangerous. For example, a fourth grade student who gets a 10th "grade-equivalent" reading score is being compared with a group of older students' reading of simple paragraphs (thought perhaps not as simple as fourth grade). However, the fourth grader COULD NOT in most cases read and understand a tenth

grade textbook, because of a lack of background knowledge and limited experience with abstract thinking. The reader who wishes to explore this quandry further is referred at this point to the topic "Grade-equivalent Scores" in the Appendix (pp.6-10). Other technical problems of reliability, equivalence, and test validity for the local district are dealt with in most college texts on educational measurement for the interested reader.16

Here, we will turn to the administrative problems of a local testing program. As previously mentioned, most norm-referenced achievement tests' scores are of limited use for judging individual students' progress. This observation implies, then, that the rationale of fall testing for individual spring program planning is not valid. Some class level plans may be made after test scores are received - to emphasize spelling or basic math more, for example. It is also important to note that in the six weeks between testing and return of scores to the district from the publisher, the teacher gives many tests, a report card, and many other kinds of work in every subject. Therefore, the tests again do not add importantly to the information she needs for planning for the spring, except in a very few cases.

Similar defects in spring testing programs are common. The May-June receipt of scores in the district means that the tests have absolutely no value for the teacher who administered them to her class; since the scores are used to place students for the fall, a student's summer experiences may limit the value of March to May testing results for the child's next teacher who can make only limited and cautious use of the tests. In both of these kinds of system-wide programs, therefore, it is clear that norm-referenced achievement tests have limited value for classroom teachers and for individual students because of the time lag in the receipt of the test results and the technical defects which limit the meaningfulness of the scores themselves. On both counts, the testing program might be said to violate a fundamental principle of activity: there has to be some clear payoff for people to get them to take the trouble to do their best.

Who, then, could be said to benefit from the common practice of taking two to four mornings some week during the year to test every child in the schools? The tests do seem to meet the needs of many parents to know how their

---

16Texts by Gronlund, Anastasi, Cronbach, Robert Thorndike and E Hagen. The definitive reference is R.L. Thorndike (ed.), Educational Measurement Second Edition,1971.

child is doing compared with his classmates or compared with the sample group of children used to "norm" the test. However, because the child's actual curriculum may not be precisely the same as that of the sample group, such comparisons are not always valuable except possibly to raise questions about the adequacy or level of the child's curriculum. Further, knowing where a child is in relation to his classmates has limited value for helping teachers to diagnose a child's specific needs or provide the necessary remedial help. However, guidance counsellors, the school administration, and the Board of Education officials can be said to have greater practical use for norm-referenced achievement test results. But even at these levels, such tests have their limitations.

At the school building level, test scores are limited in usefulness for several reasons: from building to building and class to class, curricula vary, student groups vary, testing conditions vary, and, generally, achievement test data provide less precise information even for groups of students than most test users realize.[17] Because national achievement tests are constructed to cover topics taught in schools throughout the country, it is quite possible that several questions might deal with an idea which has not been taught in a particular classroom or even in a particular district. Scores on these items will be lower for such a class. Since most achievement tests were developed before the "new" math, "new" English, etc., for a while whole sections of achievement tests were irrelevant to what some students had been learning. Information from such tests was clearly of little use to principals and school district officials in evaluating their new programs.

Student groups also vary from room to room and among schools. In some classes, students are more restless than others where students are happily quiet or forced into silence. In all of these kinds of learning atmospheres, tests will take on different meanings for the students. Happy students may hate or love tests; so may unhappy ones: test scores will reflect such factors. The conditions under which tests are given also affect scores. Some rooms are hotter than others; some teachers read directions more clearly than others; some tests are given in unusually

---

[17] Henry Dyer, "Educational testing...," N.J.E.A. Review (January, 1973), p. 32 ff. This popular article is an excellent summary of this noted expert's warnings about overreliance on tests.

long testing sessions. Again, this variety of testing situations will produce a corresponding variety of differences in scores.[18]

Finally, as is discussed in the Appendix, there is a broad allowance made by testing experts for errors: officials using tests seldom take this range of error into account sufficiently. Their most common error is believing that the tests show differences between classes when even a 5 percentile difference appears.

Because the application of nationally developed tests with simplified scoring systems has led to such problems of test misuse through misinterpretation, Henry Dyer of the Educational Testing Service recommends that test users try to examine the tests item by item, keeping in mind "the types of behavior they (are) hoping for in the children attending their schools."[19] Dyer sees hope in the development of the National Assessment of Educational Progress and of criterion-referenced tests rather than, perhaps, re-educating people to understand terms like "stanines" and "percentiles" properly.

The general dissatisfaction of test experts with national norm-referenced achievement tests has stimulated the revival of instruments like E.L. Thorndike's scales, more precise measures or "criteria" for competence in school subjects. While the search for a "precise" standard of learning is as elusive as the construction of a better mousetrap, at least the new "criterion-referenced" tests are more adaptable to the wide variety of curricula which are offered in classrooms throughout the nation.

While norm-referenced achievement tests do provide a relative picture of how much a student has learned, their scores do not really indicate what the individual actually can or cannot do. Basically, the norm-referenced achievement test score simply shows where a student stands in relation

---

[18]During some June testing, the author let one class he had been supervising take the test in an air-conditioned room; the other worked in their class. While both groups of students' work were about the same in both quality and quantity during the school year, the air-conditioned group showed a two year "grade-equivalent score" advantage over the students who had to suffer through the heat of testing in June.

[19]Op.cit., Henry Dyer, p. 35.

to other students; it puts him number 1,2,3... or 300 in his class in the subject. This does not mean that he can or cannot, say, add, subtract, or read. In fact, if you know that he actually can read, you also can say that all the students above him in score can do as well, but the main question, "can he read or not?" or "can he read satisfactorily?" is NOT answered by a norm-referenced achievement test. It only answers the question, "how well can the student read in comparison with the other students?" Therefore the major value of norm-referenced achievement tests is to classify children on the basis of their relative accomplishment. If the test is norm-referenced against a national sample, then children are classified in comparison to the accomplishments of children thousands of miles away. Therefore the popular nationally normed achievement tests have limited value for identifying specific educational deficiencies as a means of providing effective remediation. Criterion-referenced tests strive to overcome that problem.

A criterion-referenced achievement test is developed somewhat differently than a norm-referenced achievement test; for it is designed to answer the first question, "Can a student do this (whatever skill or fact learned), or can he not?" Typically, rather than collect curriculum guides like achievement test developers do, the developers of criterion-referenced tests collect information from educational research about what constitutes a competent performance. Again, much of this is being done at universities by teams of researchers, and in some areas it is easy to decide what the "criteria" of competency are. On a broader scale, the identification of "competent" reading and arithmetic performances is one of the most controversial kinds of issues.

If a "criterion-referenced" test is one which tests whether or not a student meets a particular criterion, then any well-developed test of what a teacher teaches her class is criterion-referenced. Thus, many publishers have started selling their textbook programs with "criterion" tests instead of unit tests. While there is room for some huckstering in such promotion of the old unit test under a new label, some of these new text series' tests are well developed criterion-referenced instruments. Some school systems have been able to develop their own tests over the past few years; and these, too, in some cases are well established measures of how well students have learned what they are being taught locally.[20]

Criterion-referenced tests are primarily being used in New Jersey today by schools which have individualized programs and by the State Educational Assessment Program. One easily identified group of schools moving in the direction of individualization is the League of Cooperating Schools, a national network sponsored by the Wisconsin Research and Development Center for Cognitive Learning and I/D/E/A (Institute for Development of Educational Activities, Kettering Foundation).[21] In such schools, a variety of instructional and grading patterns may be used, but in general, mathematics and reading instruction will be keyed to achieving a sequence of objectives. After each segment of the reading or math program is taught, a criterion-referenced exercise (actually a short test of ten items or so) may be done by the student. The test and other student work may then be judged by the teacher to determine what the child might learn next.

Let's say that a student has been working on breaking words into syllables at about the third grade reading level. On the criterion exercise, she succeeds in getting words like /policeman/ and /basketball/ broken up correctly but misses /camphor/ and /precise/. Since the latter group are hard words, the teacher will have to decide whether to review phonics, go ahead with syllabication, or switch to prefixes and suffixes. The judgment, ideally, will be made not on the basis of numbers but on the basis of a standard of work by a particular individual. With decisions of this sort to be made for 25 children daily, many teachers have great difficulty collecting and organizing all the information without an aide or a computer. In a school with an individualized program, general criterion tests will be given occasionally, much like achievement tests, to help

---

[20] One example of such a test series which is getting nationwid publicity is the Systematic Approach to Reading Improvemen Program, or SARI, disseminated by Phi Delta Kappa, Bloomington, Ind. The tests were developed by several school systems in Los Angeles County and Utah.

[21] Individually Guided Education and the Multiunit School, Washington, D.C.: National School Public Relations Association (1970): About 25 districts in N.J. were members in 1970. Further information may be obtained from New Careers In Education, Office of Program Development, N.J. State Dept. of Education, 1000 Spruce Street, Trenton, N.J. 08625.

the teachers decide what kinds of progress various groups of students are making and to help the principal get a general picture of the student body's learning. Computer-scored tests are being developed by many of the commercial producers of achievement tests and are already available from McGraw-Hill and Educational Development Corporation. Even at this point, it is safe to say that criterion-referenced achievement tests do provide more detailed information about students' learning needs than norm-referenced achievement tests. However, because norm-referenced tests provide limited valuable information for remediation, teachers need spend very little time analyzing them. But for teachers to utilize the information contained within criterion-referenced tests, they must spend much more time analyzing the tests to seek out the specific skills or knowledge which children have not acquired in order to provide additional help in those specific areas. As a result, some districts utilize computer assisted scoring and reporting in order to reduce the teacher's clerical time and maximize her teaching time.

IV. THE NEW JERSEY STATE EDUCATIONAL ASSESSMENT

The previous parts of this paper have suggested (1) that aptitude testing has limited utility because its scores are easily misused and because other types of information on the general student body are better for making educational decisions; (2) that norm-referenced achievement tests also have limited uses; and (3) that the newly developed criterion-referenced achievement tests show much promise for use in classrooms and schools, although they take more time to use. In this final section, the types of decisions that can be made by using the New Jersey Educational Assessment Program will be described in some detail and the problems which the tests have raised so far will be explored.

* * * *

Around 1963 when the idea of criterion-referenced testing was being revived, another group of educators developed the concept of "an education census...indicating both the progress we are making and the problems we face. This kind of information is necessary if intelligent decisions are to be made regarding the allocation of resources for education."[22] Four years were spent in developing the design and content of the project, the National Assessment of Educational Progress (NAEP, hereafter). Briefly, the NAEP covers ten subject areas, selects 9, 13, and 17 year olds in school and 17 and 26 year olds outside of school for testing, and administers a sophisticated test series nationwide. The information is carefully reported, because of political opposition from state education departments and professional associations, and does not reflect school, district, or even state level differences in performance.[23]

---

[22] Ralph W. Tyler, "Introduction," Citizenship Objectives, Ann Arbor National Assessment, p.1.

[23] "State Assessment? Many Unanswered Questions," NAEP Newsletter (September - October, 1970), p.4.

> "The NAEP sample design calls for approximately 2,000 responses to each exercise. Since the amount of time required to answer all exercises in the subject areas being assessed during a particular period is considerable (about 160 minutes in each area), not all assessees answer all exercises. To ensure that each exercise is answered by about 2,000 assessees, it is necessary to obtain a nation-wide sample size of from 25,000 to 30,000 at each of the three in-school age levels assessed.

In 1969, the NAEP was exploring state assessment design with the state of Florida. Shortly thereafter, the state of Michigan's Education Department initiated its state assessment which has set the pattern for most other states and for the controversies surrounding state assessments.[24] New Jersey's assessment came later and has its own particular history of controversy, but it avoided Michigan's pitfalls.

...If a state wished to make comparisons among schools or school districts, sampling experts estimate that a sample size of at least 500 would be required from each sampling unit in order for the data gathered to be comparable to National Assessment regional results."

It should be noted that the 1969 costs of the National Assessment test administration were about 1.8 million dollars, including scoring and printing of results. More recently, the NAEP has worked out a design for a state assessment on the National Assessment model with the state of Minnesota.

"Minnesota Pioneers 'Piggyback' Assessment," NAEP Newsletter (March-April, 1975), p. 2.

[24]Ernest R. House (ed.), School Evaluation, Berkeley: McCutchen Publishing (1973), p. 44.
"At first local districts were promised that local scores would not be made public. Generally state officials were for such reporting and school people were opposed. However, once scores showing relative standings of districts were made available to local school people, a storm of demand came from legislators for the same data. At first, the state department resisted. The fiery exchange of letters between the state superintendent and one legislator is highly amusing and informative. Like all bureaucracies under political pressure, the department finally crumbled. Local test scores were made public. Local superintendents were incensed; one charged that, 'the (program) is really politics masquerading as research.' As the authors note, assessment programs are political. As such they must serve the interests of competing groups. All in all, the Michigan assessment has been the center of more controversy than any other educational program in the state."

In New Jersey in January, 1972, Judge Botter ruled in favor of the plaintiffs in the "Robinson vs. Cahill" decision which mandated, upon review of the Supreme Court in 1973, that a system of "thorough and efficient" education as provided by the Constitution of New Jersey must be devised and financed.[25] Apparently, the New Jersey State Department of Education was well prepared for the judicial decisions, for in 1972, then Commissioner of Education Carl E. Marburger set into motion procedures for the development and administration of the New Jersey State Assessment under the direction of Dr. Gordon Ascher.[26] By June, 1972, curriculum objectives associated with test items were prepared and, as test experts recommend, checked with administrators and teachers in the state. In July and August, test items were developed by Princeton's Educational Testing Service (drawn from ETS's achievement test item bank). Test items were initially rated by teacher representatives at county meetings, on the basis of what is actually taught rather than what they thought should be taught. Subsequently, remaining potential test items were mailed to all teachers of the tested grades. Grade level committees consisting of teachers, administrators, minority representatives, curriculum specialists and students reviewed the remaining items before the tests were finally developed. Finally, a Minority Groups Advisory Council, individuals familiar with testing and minority education, recommended changes before the tests were completed. The result was essentially a statewide criterion-referenced test, in that the test only contains items on the present New Jersey curriculum and test results are reported in a way which identifies the gap between what children are presumably learning and what they actually know. The test was given for the first time to fourth and

---

[25]Gordon Ascher, "'Thorough and Efficient' and Equal Educational Opportunity: The New Jersey Mandate," Trenton: N.J. State Department of Education, Division of Research, Planning and Evaluation (1973).

[26]More information in this area is needed, but Ascher has shown that several bills were introduced into the Assembly and that Governor Cahill was persuaded to favor the program in his annual message in January, 1972, all of which suggests a long term plan was well under way before the Botter decision. (Gordon Ascher, "Utilizing Assessment Information..." (May, 1973), p. 1.)

twelfth grade pupils in November, 1972.[27] Although court challenges by the New Jersey Education Association and others held up release of the results until early 1974, the tests continued to be administered in 1973 and 1974 and they were given in 1975 again.

The New Jersey State Assessment Program consists presently of two sets of tests, one for fourth, seventh, and tenth graders which is administered every year and another for twelfth graders which is administered every three years. Each set of tests includes reading and mathematics sub-tests of about 75 and 60 items respectively. The tests are prepared for distribution from Trenton to county offices and picked up from there by school officials. On one morning in the fall, all students in the grades being tested spend about two hours taking the tests. After absentees have made up the work, all tests are returned to Trenton for scoring the tests and preparing reports which includes a detailed analysis of each child's results as well as summaries by classroom, building and district. In 1975, all districts received their individual child reports within three to six weeks of test completion, while all other reports were received after eight weeks. Along with the reports, districts receive guides for report utilization and analysis. Individual students' scores are given to the teachers and need no formal reporting to the state. Class level scores are used by teachers to prepare local reports which are collected by principals. Using the teachers' comments, each principal prepares a "Building Report" which is used by the local officials in the districts' central offices to prepare a "district summary" which has to be returned to Trenton after about 45 days. Copies of all the reports must be kept on file in local district offices for public inspection under the supervision of professionals.

While questions always may be raised about the technical quality of the tests, the central problems with the New Jersey State Assessment, as with all achievement and aptitude tests, involve the use and interpretation of the test scores at various levels. At the classroom level,

---

[27] The 1973 and 1974 tests were accompanied by some statistics on item reliability which indicated that most sections of the tests met general standards for reliability. Since the tests were developed from statewide objectives, it is reasonable to assume that some aspects of validity are also well covered by test development procedures.

teachers and parents have had to cope with a novel kind of score, the criterion-referenced score. At the school and district levels, problems have been raised by the reporting of classroom and building scores which allow (but do not require) comparisons among teachers and schools. At the district level, the program has mandated that the scores be compared with "similar" types of districts and that plans be indicated by the district for curricular change where appropriate. Overall, these interpretations are the kinds of studies which test experts have strongly recommended that all test scores deserve.

The present and potential benefits of the State Assessment Program derive from the reporting requirements as well as the criterion-referenced nature of the tests. Experts' recommendations for interpretation have commonly been treated superficially or even ignored by local district officials who are hard pressed to accomplish the basic teaching and administrative tasks of the schools. The clear priority given to evaluating students' achievement and program effectiveness is vital if the New Jersey schools are to improve over the long run. The Assessment reporting procedures make it clear that detailed and thorough thinking about students should occur in every school in the state henceforth. Such interpretation at the classroom, building and district levels will be reviewed at this point.

Recall that at the classroom level, a norm-referenced achievement test score for a student might be reported as "50" meaning "fiftieth percentile" in mathematics problem solving. The class might also get a score at the 50 percentile on a similar group of items. This might be interpreted to mean, "on a problem-solving subtest of thirty items, this student's or class's scores are the same as a group in the nationwide sample at the fiftieth percentile." Such achievement test norms do not tell the teacher whether the class got some, all, or most of the items correct, or what to teach.[28]

---

[28]Publishers' scoring services, notably Harcourt Brace Jovanovich's Stanford Reading Tests, also do some of this kind of interpretation automatically. Such data processing adds a little to the costs of standardized testing but produces much more useful data for the teachers. Again, questions must be raised about how such data is used in most cases.

However, the State Assessment scores cannot be generalized in such a fashion by the classroom teacher, a parent, or the principal. Instead, a score is reported on every single question. The idea is that each item represents a different curriculum objective within the problem-solving topic, and that to meet the criterion standard of this objective, the student must be able to get the item correct. For the class to meet the criterion standard, all the students should be able to get the item correct, in theory.

Table II illustrates some State Assessment test items at the fourth grade level.[29]

Table II. Examples of 4th Grade Test Items

---

24. How much do 4 pencils cost at 5¢ each?

(A) 5¢
(B) 9¢
(C) 20¢
(D) 45¢

33. Kim bought 2 candy bars at 10¢ each. How much change did she get from a quarter?

(A) 5¢
(B) 10¢
(C) 15¢
(D) 45¢

53. 32 is how much more than 11?

(A) 43
(B) 33
(C) 21
(D) 20

---

[29]Derived from 1973-74 and 1972 Interpretation Manuals, N.J. State Dept. of Education.

Table III illustrates how the scores on the sample items might have been reported to a classroom teacher for the 1972 tests. The teacher might interpret these scores thus: "92% of my class can figure out the cost of several items, and 96% of them can check the change they should get in the store. 85% can do a simple addition problem when it is written in the 'new math' format."

Table III Example Test Interpretation Worksheet

Item	Description of Item	relevant	Percent Correct Observed	Item Specification Comments	Cluster Comments
24	Given the cost of one pencil, tell cost of several	✓	92.5%	Satisfactory-- Some individual help indicated	Areas of satisfactory group achievement: Word problems with numbers less than 50
33	Student can tell amount of change when buying candy bars	✓	96.1%	Satisfactory	Areas for additional emphasis: Word problems with numbers greater than 50
53	Given two 2-digit numbers, tell how much more one is than the other	✓	85.7%	Small group and individual help needed	

[30]Derived from 1973-74 and 1972 Interpretation Manuals, N.J. State Dept. of Education

The teacher's judgments are indicated in the Item Specification Comments area: he or she will try to help the few kids that need work with money and to do some remedial lessons with a small group on modern math. The "cluster comments" suggest that other examples of problem solving have shown that the whole class ought to be taught how to work with larger numbers. This kind of thinking and writing had to be done for the 75 or so reading and 60 math examples.

This example gives a rough indication of the main professional problems the New Jersey State Assessment test scores produced; for every fourth, seventh, tenth, and twelfth grade math teacher has, in the past three year testing cycle, examined a considerable portion of the same test, example by example, after receiving a little instruction on how to read the scores. Few school systems had devoted such time to interpreting tests previously because the test publishers' scores were not so detailed and because classroom testing often covered the same ground.[31]

When scores were made available, classroom by classroom, moreover, teachers experienced judgments about their classes' performance in comparison to other teachers' classes even though classes weren't identified by the state forms. The "political" effects of teachers' general awareness of student success can hardly have been admitted openly, but it must have contributed to the negative reaction to the state testing program. Thus, at the classroom level the New Jersey State Assessment Program produced some potentially useful information which was not previously available to teachers. At the same time, the professional, economic, and political factors which accompanied this interpretation process, may have made the information less used in practice than it might have been if controversies hadn't arisen. It may take some time for this type of reaction to disappear, but, again, in the long run, the scrutiny of achievement by the professionals at the local level itself will produce pressure for improvement. Thus, the State Assessments provide an opportunity for the greater professionalization of the educational field.

---

[31]The best teachers have often reviewed tests and scores thoroughly, when they had time; and most school systems hold periodic workshops on achievement and aptitude test interpretation. Again, nothing like the detailed written reports of the State Assessment were produced in such individual or system-wide reviews of test scores since the 1920's.

Table IV Examples of Comparative Test Interpretation Worksheet

ITEM SPECIFICATIONS AND DATA			PHASE I INTERPRETATION	PHASE II
Item number and description (Record all items) (2)	Relevant	Percent Correct Observed (3)	Item Specification Comments (4)	Cluster Comments (5)	Community Type
Given the cost of one pencil the student can determine cost of several pencils (less than 10)	✓	92.5%	Satisfactory some individual help indicated	Areas of satisfactory group achievement/some individual or small group remediation	Relevant 91.0
Student can determine amount of change he should receive from a quarter after buying 2 candy bars	✓	91.6%	Satisfactory	Word problems with numbers less than 50 (items 23, 33, 52)	Relevant 87.5
Given two 2-digit numbers the student can determine how many more one is than the other	✓	85.7%	Small group and individual help needed	Areas for additional emphasis	Relevant 65.7

PHASE II INTERPRETATION

ɔnal	County	State	Item Specification Comments	Cluster Comments
	92.0	95.1	Roughly equivalent to reference groups.	For the problem solving cluster, emphasis is needed on word problems with larger numbers even though the areas needing attention were low for the reference groups.
	89.6	94.4	Greater achievement than reference	Satisfactory achievement is noted for problem solving with larger numbers and this area is slightly higher than the reference groups.
	67.2	68.7	Initially diagnosed as need area. However, greater achievement than reference.	

rived from 1972, 1973 and '74 Interpretation Manuals, New Jersey State Dept. of Education.

The school principals also filled out a second set of forms comparing their schools to the district-wide scores. At the district level, similar comparisons were made to various types of school districts state-wide. Table IV shows some sample comments at this level of interpretation. Under "community type" the figures indicate scores for communities classified as one of ten urban, suburban, or rural categories.[33]

The Table IV data are quite different in format from the district-wide reports traditionally received from commercial achievement testing, and professional problems in Assessment reporting similar to those raised by teachers may be assumed to have occurred. Principals and curriculum workers had to be trained in test interpretation in order to train the teachers. They also had to write the voluminous reports on the tests as well as narrative reports for the state. The state reports, moreover, requested that administrators "suggest appropriate means for alleviation of needs" which the tests revealed.[34] This has led both to internal and external problems. For example, while the teacher who completed Table III might have thought her class needed more work on modern math, because the class results were lower than the standards she or others had expected of them, the comparative scores (which were not usually available to teachers) show that the class as a whole is generally mastering mathematics better than other groups of students in New Jersey. Would the possession of that information tend to deter the teacher from carrying out her planned remedial activities? What if her class' performance were below the referent groups? How would her thinking change? For that matter, how do administrators view such comparative data in evaluating building and district achievement and what effect does it have on educational planning? The answers to these questions are crucial to maximize the value of the assessment process.

---

[33]"Test Interpretation," Office of State Assessment, Trenton: N.J. Department of Education (1973-74), p.20.

[34]Ibid., p. 11.

In some affluent districts, there has been concern that students did not do as exceptionally well as some persons anticipated. The data, despite suggestions by the Assessment officials, were used by newsmen and citizens to make comparisons among districts. Internally, there is always the possibility that teachers and administrators in schools with scores below the district "average" received undue pressure to find ways to raise their students' performances by mechanistically teaching only to the test rather than making effective program improvement and, conversely, that officials of better schools reflected some complacency unwarranted by test results. Some schools began long term improvements which will not bear fruit for several years. Finally, the testing itself has been politically and economically expensive to interpret. In some districts educational association pressures have led to the hiring of numbers of substitutes so that teachers could meet, instead of teaching, to work on interpreting the test scores. Statewide, the program has come under attack as overly expensive although its per pupil costs are low in comparison to amounts usually budgeted for testing locally.

Because the State Assessment Program was the first aspect of the Thorough and Efficient plan to emerge after the Robinson vs. Cahill decision, it has also to some extent borne the brunt of the N.J.E.A.'s opposition to a perceived threat of state uniformity in local education as well as political opposition to a perceived threat of increasing state control.[35] It is hard to imagine how the State Department of Education could have begun to execute its mission to provide equal educational opportunity in New Jersey without arousing some opposition. If the local schools had been accustomed to a little more state activity, as is the case in New York and Pennsylvania (where similar systems have been introduced more gradually and with less controversy so far), the assessment program up to the summer of 1975 might have developed in a less heated atmosphere.

---

[35]One fear of the N.J.E.A. was that assessment scores would be used to rate teachers, as did occur in Michigan early in its Accountability Program. Such misuse of test data, however, occurs also with standardized achievement tests, and this writer has no evidence that such abuses of test information have occurred in New Jersey because of the Assessment Program.

The main point, then, is that the State Assessment Program is a key element of "T&E" and as such is here to stay. It is a definite improvement upon the variety of commercial programs now used throughout the state, for the Assessment Program is more relevant, uses less student classroom time, and produces better decision-making data at all levels than the commercial norm-referenced achievement tests usually do. It helps the classroom teacher to identify the specific areas in which children need help. As part of a district self-evaluation process, it provides administrators and board members with invaluable data to help assess progress towards planned goals. Planning can be improved by the possession of information regarding the performance of other districts on identical questions, information which was never previously available due to the variety of commercially prepared achievement tests used throughout New Jersey. The Assessment Program has already provided considerable data to help the public learn where its school programs really stand both in terms of local goals and by comparison with others. It promises to provide much needed assistance to all New Jersey educators and citizens in their cooperative efforts to make realistic and continuing improvements in the schools.

GLOSSARY OF TERMS

Compentency Indicator -- A measurable or observable behaviour or variable used to determine the extent of a student's skill or knowledge.

Intelligence -- A general aptitude for learning psychologically defined in several ways as follows:
1) Performance on an "intelligence" test -- (this is a very technical definition, for you must know about intelligence testing to understand its implications.)
2) "The summation of the learning experiences of the individual".[35] (A general definition)
3) "The aggregate capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment." (Wechsler, op.cit.) (A very broad definition)

I.Q. -- "Intelligence Quotient." A simplified score which is the ratio between an intelligence test score called mental age and the individual's chronological age; now considered by most psychologists to be so grossly misinterpreted that it is very unpopular for group test interpretation. Computed as follows:

$$\frac{\text{Mental Age Score}}{\text{Chronological Age}} \times 100 = \text{IQ}$$

Score -- Any numerical representation of an individual's performance on a measure of some behavior. May include ratings of work, learning, physique, etc. Types of scores are discussed in the guide below.

Skill -- 1. "degree of mastery already acquired in an activity."[36] This definition might be more clearly expressed as "skill level", for otherwise it is easily confused with 2.
2. the process of performing an action. Complex actions such as reading and computing can be broken down for teaching purposes into specific "skills," even though the process itself, when learned, is performed smoothly and without apparent recourse to step-by-step "skills."

---

[35] L. Wendell Rivers et al., "I.Q. Labels and Liability," St. Louis, Mo., (unpublished position paper), n.d.

[36] John O. Crites, Vocational Psychology, New York: McGraw-Hill Book Co. (1969), P. 26.

Standard -- a stated level of attainment which defines satisfactory competency. It may be expressed in terms of test scores or other quantitative measures.

Technical Testing Terms -- Various ideas connected with designing, developing, and proving the validity of tests, types of scores, and population sampling are some of the technical aspects of testing mentioned in the guide below. The interested reader, however, should refer to an introductory textbook on educational testing, such as Thorndike-Hagen or Cronbach.

1) An objective test is one which is taken and graded under virtually identical conditions Multiple choice questions are considered objective as opposed to essay questions which are considered subjective.
2) A test is reliable to the extent that it accurately measures the child's performance. Accuracy may be analyzed as consistency in performance from the beginning to end of one test or as consistency of test scores over a time lapse of a few weeks or months.

Text Experts -- Groups of university researchers and of publishers engaged in designing, developing, and marketing tests as well as in theorizing about educational measurement. They are identified most usefully for the practitioner in Oscar Buros (ed.), Mental Measurements Yearbook, Highland Park, N.J.: Gryphon Press, (1972), which indicates that of the 1,157 tests in print, the following publishers control about 40% of the market.

Publisher	Number of Tests Published
Educational Testing Service	96
College Entrance Examination Board	87
Harcourt-Brace-Jovanovich (HBJ)	63
Psychological Corporation	47
Cooperative Testing Corp.	46
Western Psychological Services	46
Data Processing and Educational Measurements Center	40
California Test Bureau/McGraw-Hill	30
Houghton-Mifflin	.25

The E.T.S. and Harcourt-Brace conglomerates alone publish about 350 of these tests, or nearly 30% of the market, and recently HBJ announced its teasting department would incorporate the Psychological Corporation.

Test Types -- 1. Informal: any test prepared to evaluate anything -- most classroom testing.

2. Standardized: a test which all students take under similar conditions, as regards procedures, test materials and scoring. Test materials have been subjected to technical development and validation procedures, hence to some degree meet standard quality criteria.

3. Norm-referenced: a test which contains questions of varying difficulty so that the test results of a representative sample group fit a normal curve (i.e., half the group will score above and half will score below the average score of the group). Test results are generally used to classify individuals in accordance with their relative scores.

4. Criterion-referenced: A test which contains questions which competent students are presumed to be able to answer correctly. Test results are used to identify the gap between what children are expected to know and what they actually know so that the teacher may concentrate her efforts where needed most.

5. Aptitude: (includes intelligence) a test which reflects the cumulative effect upon an individual of both school and non-school general learning experiences and is frequently used to predict future performance.

6. Achievement: a test which measures what an individual knows or can do as a result of specific instruction or training and is used to assign grades, diagnose learner needs in order to prescribe remediation programs, evaluate and improve teaching, and formulate educational goals.

7. Battery: a group of tests from a single publisher used to test several aspects of ability or achievement. The individual tests in the group are called "sub-tests."

Examples: Stanford-Binet Intelligence Scales
Wechsler Intelligence Scales for Children and for Adults (WISC, WAIS)
These are both administered by qualified experts, to one person at a time.

California Achievement Tests, Metropolitan Achievement Tests, Iowa Tests of Basic Skills, Stanford Achievement Tests, etc. Often the reading or math subtests are

singled out by publishers and re-labeled as specific achievement tests.

Differential Aptitude Tests, General Aptitude Test Battery (U.S. Employment Service). These are two of the best known batteries of aptitude tests for group administration.

8. Quantitative and non-verbal Subtests: sections of aptitude or intelligence tests that do not require answers in terms of words. Quantitative tests deal with mathematical problems; non-verbal tests may deal with visual puzzles also.

9. Performance Tests: do not require paper and pencil. The Red Cross swimming tests, on-the-road drivers' licensing tests, and airlines' physical examinations and reaction-time tests for pilots are examples of some types of performance tests. Other types may involve reading directions and then assembling equipment.

A GUIDE TO TEST SCORES

The problem of using achievement tests well is compounded by the variety of scores which can be produced statistically these days. Rather than list such terms, it will be helpful for the general reader to get a picture of what happens as someone scores a group of tests and to read the technical terms in context. For more technical definitions, the interested reader is again referred to standard texts on educational measurement, such as ones by Thorndike-Hagen, Cronbach, or Anastasi.

Testing situation: for the purposes of discussion, let us take a highly simplified test situation. Instead of a K-8 school system of 1000 students, let's consider four sixth grade children as the "class," and rather than a whole test battery, only one subtest. Scoring procedures would be identical for longer tests and larger student groups. Test interpretation, of course, gets more complex and time consuming as the test size and student group increase.

So, the 4 students have had twenty minutes to answer 40 questions as the teacher has administered the Stanford Achievement Test's Vocabulary Subtest by reading incomplete sentences aloud. The students then mark one of four choices on their papers to complete the sentence. If these were hand-scored rather than being sent to a computer, the teachers would then go through the scoring steps and a variety of interpretive procedures.

SCORING

Step 1: RAW SCORES: The number right is totalled for each student. This is called a "raw score." Annie Ames and Billy Burke each answer 20 items or half the test correctly. Carol Chaim got 10 right; Denny Darn got 30 right. In terms of simple percentages (number right divided by total number of questions), this means:

A scored 50% (20 right) C scored 25% (10 right)
B scored 50% (20 right) D scored 75% (30 right)

This test was deliberately designed so that an "average" score involves getting most items right.

Step 2: SCALED SCORES: Next, the teacher refers to a scoring table in a Teachers Manual for the test. This scoring table was prepared from the pilot testing of a national group, called a NORM SAMPLE, and the scores all refer to the NORMS of the national group. Thus, the tests are often described as NORM-REFERENCED. Commercial test publishers can develop LOCAL NORMS after a test has been administered in a school district for a couple of years. These LOCAL NORMS, however, will not be printed in a test manual.

Three types of commonly used SCALED SCORES will be shown: PERCENTILES, GRADE-EQUIVALENT SCORES, and STANINES. Other types of scores are Z-Scores, T-Scores, and Standard Deviations. The College Boards use a special type of score derived from Standard Deviations.

Other types of tests produce scores by comparing a student's performance to a CRITERION of competence. These CRITERION-REFERENCED scores will be discussed below and have been extensively discussed in Section IV of this paper.

The first kind of score, the one which seems to show the students' relation to the national norm group most clearly, is a PERCENTILE. The scores below for each of the children tested above are percentiles taken from the Stanford Reading Test manual for teachers.[37]

A - 10 %ile	C - 1 %ile
B - 10 %ile	D - 50 %ile

Once the teacher has such figures for her students, she can begin to INTERPRET the scores, that is think about what the numbers mean in terms of students' learning programs. These figures might suggest that although all of her sixth graders succeeded in answering at least ten of forty questions correctly on the test, only about 1% of the students in the nation would get such low scores. This might lead the teacher to identify all the students falling below the 10%ile level as needing special help in reading. For her purposes, it doesn't matter how she arrived at the scores. Since the scores are based only on number right, it makes no difference whether the student answered easy or hard questions. The percentiles come out the same either way. In other scoring systems, commonly used on intelligence tests, individually given reading diagnostic tests, and some other group tests, scores are prepared by counting number of consecutive correct answers until a "ceiling" is reached. In the class being discussed, where simple raw scores are used, the low percentiles might hide the fact that even the student who scores at the 1%ile level CAN READ something. His or her competence is being compared to that of similar aged students, however, not to a CRITERION of reading competence.

---

[37] Bjorn Karlse, et al., Manual for Administering and Scoring Stanford Diagnostic Reading Test Level II (1966). New York: Harcourt Brace & World, p.32.
Note that the student with 75% right is only at the 50 percentile nationally because of the test design discussed above.

Let us return to the problem of interpreting scaled scores. The teacher might have chosen to use a different scoring table to get GRADE-EQUIVALENT SCORES. If the students had taken the whole test and gotten about the same proportions of correct scores, a different table would have shown the GRADE-EQUIVALENTS to be:

A - 4.6	C - 2.4
B - 4.6	D - 7.5

The Manual explains these scores very superficially. They are produced through a complex process of matching students' scores on particular items to those of older students. Because few such items are used in the increasingly harder levels of the test, this matching is a very crude estimation. For instance, questions 25-30 of the Level II test might be like questions 1-5 of Level III in difficulty.[38] A score derived from computing young students' answers and older students is only an estimate of how each group did on the rest of their respective tests. The Level II test is likely to be much easier than Level III throughout, and it would be hard to say that a 7.5 Level II (30 correct) is equivalent to a 7.5 on Level III (perhaps 10 correct items) when the rest of the Level III form of the achievement test is quite different in subject matter and difficulty from Level II. Thus, the grade-equivalent scores are subject to much the same kind of misinterpretation that plagues the I.Q. concept.[39] A sixth grade student who gets a 10.5 on the Stanford Reading Test, Level II form could probably NOT read and understand any tenth grade textbook because he has not got enought background knowledge and ability to think using abstractions for such difficult reading.[40]

If the above grade-equivalent scores were INTERPRETED by the teacher, without the benefit of percentiles, she might judge that only Carol, student C, needed help in reading because Carol is more than two years below grade level. Clearly, this will mean less reading instruction will be planned for students A and B because, according to these scores, they are not very far below the sixth grade level. Clearly, it is most important to have the teacher

---

[38] See op.cit. Lyman pp. 115-118 or op.cit. Gronlund pp.374-79 for more detailed discussion of these and other technical points.

[39] Op.cit. Gronlund, p.376. He cites recommendations by test experts that grade equivalent scores be dropped.

[40] Ibid.,p.378.

making judgments using the scores which will benefit her students most, and in keeping with the goal of re-educating the professionals to make better judgments, the Stanford Manual tries to de-emphasize the grade-equivalent scores.

But what, then, is the right answer about this class's scores? Are two students in trouble in reading who are not going to receive the attention they need? If the teacher follows the Manual further, the situation gets a little more complicated because the Manual for the Stanford Reading Test warns that the tests produce much less accurate information about pupils that about classes: " ...only differences of two or more stanines for the results of an individual pupil should be considered meaningful. For class performance, a difference of one half stanine between two subtests is generally significant." In order to make an accurate judgment, therefore, the teacher should refer to a third type of score, the STANINE, which is closely related to percentiles. This table illustrates how the two types of scores are related.

Stanine	1	2	3	4	5	6	7	8	9
Percentile	4	11	23	40	60	77	89	96	

In other words, if there has to be a TWO STANINE difference between two pupils for it to be considered valid, there is NO difference between students A, B, and C, according to the subtests of the Stanford Reading Test. The teacher should, therefore, study these students much more closely before making any judgment about their needs for special help in reading. Such study may include giving them some individual work to diagnose the problems they had on the test, or it may just involve giving them another kind of test.

The reason that such a large difference in scores is needed for individuals is that many different factors can affect different people in a group. Therefore, a margin for error in testing is computed, the STANDARD ERROR OF MEASUREMENT, or STANDARD ERROR for short. When a group of individuals' scores are combined, producing a CLASS SCORE these individual errors tend to cancel each other out, and therefore the standard error for a group is much smaller than that for an individual. Thus, half a stanine difference in classes' scores may tell the administration of a school district a good deal about which classes need most help for reading specialists.

Because of the difficulty in properly interpreting scores and because of the standard error of measurement for individual scores, the commercial norm referenced achievement

tests commonly given in New York and New Jersey are useful mostly outside of the classroom. This implies that the parents, seeing their individual child's scores, and teachers, looking only at their students' scores or their single class's scores, don't benefit much from the hours of work and tension that divert the children from regular instruction in the periods of time set aside for mass achievement testing. The main result of several hours of general testing is that the teacher learns who was at the top or bottom of the class in various areas, facts which she learns faster on her own, by just teaching and observing her students at work.

A CRITERION-REFERENCED test on vocabulary would be quite different from the Stanford Reading Test. On the Stanford, it didn't matter which items Ann, Bill, Carol and Denny got right. Their scores would come out the same no matter which ones were wrong. On a criterion referenced test, however, the vocabulary scores might be broken down in reference to various types of vocabulary knowledge criteria. The following are hypothetical examples of such a breakdown and the score reports that might accompany it:

Criterion Topics	Ann	Bill	Carol	Denny
10 sight words (synonyms)	8	5	0	10
10 words in context	7	10	5	10
10 multi-syllable words	5	3	5	8
10 abstract words	0	2	0	2

Clearly, this type of test score is much more complex to produce and to interpret. One advantage of a CRITERION-REFERENCED score is that generalizations are not made to hide the differences among pupils: the total vocabulary score in this case doesn't mean as much as the scores on the separate criteria. Another advantage is that, when the teacher takes the time to study the scores, it is very clear which students need which kinds of reading lessons. Bill and Carol need a lot of work on synonyms. Everybody needs to learn abstract words sooner or later. Three of the children should have a good deal of work with multi-syllabic words, and it is interesting that in this harder area, Carol seems to be doing about as well as anyone. The teacher might perceive this fact as a clue for further study of Carol's instructional program or of her test-taking behavior. Perhaps she just missed the first few questions because the boy next to her was whispering! Hopefully, this little excursion makes it clear to the reader why criterion-referenced tests are difficult to use but have potential for improving the educational practices in schools.